Computers Math. Applic. Vol. 24, No. 8/9, pp. 1–16, 1992
Printed in Great Britain.

0097-4943/92 $5.00 + 0.00

SYSTEMS OF SEEMINGLY UNRELATED REGRESSION EQUATIONS WITH TIME VARYING COEFFICIENTS— AN INTERPLAY OF KALMAN FILTERING, SCORING, EM- AND MINQUE-METHOD

WOLFGANG SCHNEIDER[†]
Institute of Statistics and Econometrics
Faculty of the Social Sciences, University of Kiel
Kiel, Germany

Abstract—This paper gives a detailed description of the implementation of ML-estimation using scoring and EM for the hyperparameters of a particular econometric state space. We also show how the EM-method can be turned into an on-line estimation method and how this procedure relates to an adaptive method of variance component estimation based on MINQUE-theory. A small simulation study for a single equation model indicates how much can be gained in terms of computer time by using a combination of scoring and EM.

INTRODUCTION

The organization of the paper is as follows: First we cast a system of seemingly unrelated regression equations with time varying parameters into the framework of an econometric state space model. We then apply Kalman filtering to this model in order to calculate estimates of the time varying parameters (the states) and to calculate the likelihood function for its hyperparameters. In addition we present a fully recursive square-root filter for combined filtering and smoothing and give a simple updating formula for the smoothed cross products of the states. These procedures should enhance and simplify known implementations of the EM-method in the literature as, e.g., in [1].

Starting from an elementary exposition of scoring and the EM-method, we then go on to show in detail how these methods may be applied to the estimation of the hyperparameters of the econometric state-space model considered here. The EM-estimators turn out to have a familiar AITKEN structure involving the smoothed states and their crossproducts. Approximating the smoothing solution by the corresponding filtering solution allows us to turn the EM-method into a fully recursive estimation method for the hyperparameters ("adaptive filtering" as it is called in the engineering literature). This filter will considerably reduce the storage requirements for intermediate results as compared to the usual EM-procedure. The following section points out the close relationship between the approximate EM-method and an adaptive filter proposed by Louv [2], who applied ideas from variance component estimation (MINQUE-theory) to the estimation of the noise variances in a state space model. A simulation study for a single equation model concludes the paper, demonstrating that a lot can be gained in terms of likelihood increase per computer time unit by using scoring in combination with one of the EM methods. The latter generate large, computationally cheap likelihood increases during their first iteration steps and produce a convenient starting point for a final step of scoring.

[†]The author was killed in an accident in India on October 16, 1990.

SETTING UP THE ECONOMETRIC STATE SPACE MODEL

We will consider the following system of seemingly unrelated regression equations:

$$y_t = U_{1t}'x_t + U_{2t}'\beta + v_t, \tag{1a}$$

$$x_t = \Phi x_{t-1} + U_{3t}'\gamma + w_t, \tag{1b}$$

where y_t is an $(m \times 1)$-vector of observables to be explained as a linear function of observables U_{it} $(i = 1,2)$ superimposed by white noise $\{v_t\}$; the coefficient vector $x_t(k \times 1)$ of U_{1t} are time varying, whereas the coefficients $\beta(l \times 1)$ of U_{2t} do not vary over time. The behaviour of x_t obeys the law of motion (1b), which is a first order autoregressive scheme superimposed by some systematic influence U_{3t} and driven by white noise process $\{w_t\}$. (Also higher-order dynamics may be reduced to this parametric setup, see, e.g., [3].) The matrices U_{it} are blockdiagonal, where each diagonal contains a row vector of the explanatory variables corresponding to that equation. Collecting all explanatory variables in the $(n \times 1)$-column vector u_t we have:

$$U_{it}' = (I \otimes u_t')C_i \quad (i = 1,2,3),$$

where the C_i's are appropriately chosen selection matrices. The coefficient vectors x_t, β, and γ are partioned conformably. Alternatively, we may express system (1) as

$$y_t = U_{1t}'x_t + Bu_t + v_t, \tag{2a}$$

$$x_t = \Phi x_{t-1} + \Gamma u_t + w_t, \tag{2b}$$

where

$$\operatorname{vec}(B') = C_2\beta, \quad \operatorname{vec}(\Gamma') = C_3\gamma. \tag{2c}$$

In a completely analogous way, we collect the nonzero elements of Φ in φ:

$$\operatorname{vec}(\Phi') = C_4\varphi, \quad \text{i.e.,} \quad \Phi x_t = (I \otimes x_t')C_4\varphi. \tag{2d}$$

Matrices C_i's may also be interpreted to embody arbitrary linear restrictions on the parameters Φ, B, and Γ. Observables are available over the time span $t = 1,2,\ldots,N$; the variables in u_{it} are taken to be nonstochastic. (We note that under Gaussian assumptions we can also allow lagged dependent variables among the u_{it} and other stochastic regressors independent of all random variables in (3) and still preserving the linear structure of Kalman filtering; for a thorough discussion of stochastic regressors in state-space filtering see [4,5].) The stochastic properties of $\{y_t\}$ will be derived from the joint distribution of the random vectors $\{x_0, w_0, \ldots, w_N, v_1, \ldots, v_N\}$.

ASSUMPTION. *The vectors* $\{x_0, w_1, \ldots, w_N, v_1, \ldots, v_N\}$ *are mutually uncorrelated and form a multivariate normal distribution, where*

$$x_0 \sim \mathcal{N}(\mu_0, \Sigma_0), \quad v_i \sim \mathcal{N}(0,R), \quad w_i \sim \mathcal{N}(0,Q), \quad i = 0,1,\ldots \tag{3}$$

Equation (1) and the stochastics (3) specify a particular econometric state space model, where (1a) is the measurement equation and (1b) being the transition equation of the states x_t.

Stated in Bayesian terms: given the system parameters $(\Phi, \gamma, Q, \mu_0, \Sigma_0)$ the transition law (1b) and assumption (3) specify a prior distribution on the time-varying coefficients $\{x_t : t = 0,1,\ldots\}$; whereas the complete system (1) yields in conjunction with (3) a likelihood function
$f(y_1, \ldots, y_N; \theta)$ for the whole set of hyperparameters $\theta := (\beta, \gamma, \Phi, Q, R; \mu_0, \Sigma_0)$. (The Bayesian viewpoint of state space modelling is taken, e.g., in [6–8].) The purpose of the statistical analysis of (1) is to solve the following four problems:

(a) Reconstruction of the historical and future path of the time-varying coefficients $(x_t : t = 0,1,\ldots)$ by their conditional means $\{E(x_t \mid y_1, \ldots, y_N; \theta)\}$. Given a quadratic loss function these means are optimal Bayes estimators of x_t based upon the available sample

information $y(N) := \{y_1, \ldots, y_N\}$; depending on whether $t < N$, $t = N$ or $t > N$ they are called *smoothing, filtering*, or *prediction solution*, respectively.

(b) Calculation of a measure of precision for the estimated parameter path, e.g., the covariances $\{\mathrm{COV}(x_t \mid y_1, \ldots, y_N; \theta)\}$.

(c) Calculation of an estimate $\hat{\theta}$ for the unknown hyperparameters θ according to the ML principle.

(d) Derivation of an asymptotic measure of precision for the estimated model parameters, e.g., by an appropriate description of the likelihood curvature.

KALMAN RECURSIONS

The Kalman filter algorithm provides a convenient instrument for solving problems (a) and (b). The filter permits one to recursively calculate arbitrary *a posteriori* distributions for the states x_t given sample information $y(s) := \{y_1, \ldots, y_s\}$ along with predictive distributions for the observables y_t given sample information $y(t-1)$. Under the above assumptions these distributions are normal with:

$$\hat{x}_{t|s} := E(x_t \mid y(s); \theta), \qquad \Sigma_{t|s} := \mathrm{COV}(x_t \mid y(s); \theta), \tag{4a}$$

$$\hat{y}_t := E(y_t \mid y(t-1); \theta), \quad D_t := \mathrm{COV}(y_t \mid y(t-1); \theta). \tag{4b}$$

Smoothing solutions ($s = N$) are computed via a series of forward and backward recursions (see [9])

initialization:

$$\hat{x}_{0|0} := \mu_0 \quad \Sigma_{0|0} := \Sigma_0, \tag{5}$$

forward recursions:

$$\hat{x}_{t|t-1} = \Phi \hat{x}_{t-1|t-1} + \Gamma u_{3t}, \qquad \Sigma_{t|t-1} = \Phi \Sigma_{t-1|t-1} \Phi' + Q, \tag{6a}$$

$$\hat{y}_t = U_{1t}' \hat{x}_{t|t-1} + U_{2t}' \beta, \qquad D_t = U_{1t}' \Sigma_{t|t-1} U_{1t} + R, \tag{6b}$$

$$\hat{x}_{t|t} = \hat{x}_{t|t-1} + K_t (y_t - \hat{y}_t), \qquad \Sigma_{t|t} = (I - K_t U_{1t}') \Sigma_{t|t-1}, \tag{6c}$$

backward recursions:

$$\hat{x}_{t|N} = \hat{x}_{t|t} + A_t (\hat{x}_{t+1|N} - \hat{x}_{t+1|t}), \qquad \Sigma_{t|N} = \Sigma_{t|t} - A_t (\Sigma_{t+1|t} - \Sigma_{t+1|N}) A_t', \tag{7}$$

where

$$K_t := \Sigma_{t|t-1} U_{1t} D_t^{-1}, \quad \text{Kalman filter gain,} \tag{8a}$$

$$A_t := \Sigma_{t|t} \Phi' \Sigma_{t+1|t}^{-1}, \quad \text{Kalman smoother gain.} \tag{8b}$$

ENHANCEMENT OF NUMERICAL PRECISION

Recursions (6)–(8) are known as the *standard covariance form* of the Kalman filter. A disadvantage of this numerical variant is the fact that taking differences in (7) and (6c) might produce rounding errors which possibly accumulate as to render the computed covariance matrices no longer positive semi-definite. This problem may be avoided by using the so-called square-root filter [10,11]. There are fully recursive square-root versions of the *combined* filter and fixed-interval smoothing recursions for the so-called information filter [12–14], which processes the inverse of the covariance matrices $\Sigma_{t|s}$. A version of combined filtering and fixed interval smoothing for the standard covariance filter—a solution arrived at by a simple extension (see [15]) of the usual square-root filtering formulas—is given below. In the sequel, we assume all covariance matrices to be positive definite and denote their lower-triangular Cholesky root by a tilde above the matrix.

Recursions of the Square Root Filter

forward filtering:

$$\begin{bmatrix} \tilde{\Sigma}'_{t|t-1} & \tilde{\Sigma}'_{t|t-1}A'_{t-1} \\ 0 & \tilde{B}'_{t-1} \end{bmatrix} = \mathbf{H}_1 \begin{bmatrix} \tilde{Q}' & 0 \\ \tilde{\Sigma}_{t-1|t-1}\Phi' & \tilde{\Sigma}'_{t-1|t-1} \end{bmatrix}, \tag{9a}$$

$$\begin{bmatrix} \tilde{D}'_t & \tilde{D}'_t K'_t \\ 0 & \tilde{\Sigma}'_{t|t} \end{bmatrix} = \mathbf{H}_2 \begin{bmatrix} \tilde{R}' & 0 \\ \tilde{\Sigma}'_{t|t-1}U_{1t} & \tilde{\Sigma}'_{t|t-1} \end{bmatrix}, \tag{9b}$$

backward smoothing:

$$\begin{bmatrix} \tilde{\Sigma}'_{t|N} \\ 0 \end{bmatrix} = \mathbf{H}_3 \begin{bmatrix} \tilde{\Sigma}'_{t+1|N}A'_t \\ \tilde{B}'_t \end{bmatrix}. \tag{9c}$$

The matrices $\mathbf{H}_i$'s are orthonormal Householder transformation matrices which triangularize the matrix premultiplied by the transformation matrix. The matrices A_t and B_t are auxiliary quantities, A_t being already defined in (8b). B_t may be identified as a component of recursion (7). We have:

$$B_t = \Sigma_{t|t} - A_t\Sigma_{t+1|t}A'_t = \Sigma_{t|t} - \Sigma_{t|t}\Phi'\Sigma^{-1}_{t+1|t}\Phi\Sigma_{t|t}.$$

The recursion may be verified by taking the square on both sides of (9) and thus arriving at the original recursions of the standard form.

CALCULATION OF SMOOTHED CROSSPRODUCTS OF THE STATE

For ML-estimation of (1), we also need the smoothing solution for the crossproducts $\Sigma_{t-1,t|N}$ of successive states, where

$$\Sigma_{t-1,t|N} := \mathrm{COV}(x_{t-1}, x_t \mid y(N); \theta). \tag{10}$$

One way to do this is to expand the state as to include x_{t-1} along with x_t as, e.g., suggested by Watson and Engle [1, p. 395]. This, however, unnecessarily blows up the dimension of the filtering problem, since there is a very simple relationship between the cross-moments (10) and the filtering solution at the horizon N. We have:

$$\Sigma_{t,t+1|N} = A_t\Sigma_{t+1|N}. \tag{11}$$

An easy proof using projection arguments in the spirit of [16] would proceed along the following lines. From the smoothing recursions (7), we derive the following difference equation for the errors $\tilde{x}_{t|s} := x_t - \hat{x}_{t|s}$:

$$\tilde{x}_{t|N} = \tilde{x}_{t|t} + A_t(\tilde{x}_{t+1|N} - \tilde{x}_{t+1|t}). \tag{12}$$

Postmultiply (11) by x_{t+1} on both sides, recalling that x_{t+1} may be decomposed into

$$x_{t+1} = \hat{x}_{t+1|t} + \tilde{x}_{t+1|t} \quad \text{or} \quad x_{t+1} = \hat{x}_{t+1|N} + \tilde{x}_{t+1|N},$$

then take expected values on both sides. Observing that $\tilde{x}_{t|s}$ has zero expectation given $y(s)$ one can deduce result (11). This result may also be inferred from [17] and [18], who use a more complicated approach.

SETTING UP THE LIKELIHOOD FUNCTION

The Kalman filter recursions (6) provide us with the parameters of the conditional normal densities $f(y_t \mid y(t-1); \theta)$. Hence, the log-likelihood for θ may be recursively calculated (except for a constant term) as:

$$L_N(\theta) = -\frac{1}{2}\sum_{t=1}^{N}(\log(|D_t(\theta)|) + [y_t - \hat{y}_t(\theta)]'[D_t(\theta)]^{-1}[y_t - \hat{y}_t(\theta)]). \tag{13}$$

This function is to be maximized with respect to θ. We assume that the solution can be found by differentiation within the interior of the parameter space, i.e., the solution, $\hat{\theta}_N$, satisfies $\nabla L_N(\hat{\theta}_N) = 0$ and $\nabla^2 L_N(\hat{\theta}_N) < 0$.

SCORING

The scoring method [19] for the computation of $\hat{\theta}_N$ is an iterative search procedure consisting of the following iterations:

$$\hat{\theta}_N^{(i+1)} = \hat{\theta}_N^{(i)} + \alpha[\hat{\mathbf{I}}_N(\hat{\theta}_N^{(i)})]^{-1}\nabla L_N(\hat{\theta}_N^{(i)}). \tag{14}$$

Parameter α denotes an appropriate step length (optimized as, e.g., in [20, pp. 11–13]) and $[\hat{\mathbf{I}}_N(\hat{\theta}_N^{(i)})]$ an estimate of the information matrix $\mathbf{I}_N(\theta) := E[-\nabla^2 L_N(\theta)]$ evaluated at $\theta = \hat{\theta}_N^{(i)}$.

Following the rules of matrix differentiation (see, e.g., [21]), we compute the derivative of (14) with respect to the i^{th} component of θ as:

$$\nabla_i L_N(\theta) = -\frac{1}{2}\sum_{t=1}^{N}[\text{trace}\{D_t^{-1}\nabla_i D_t(I - D_t^{-1}\tilde{y}_t\tilde{y}_t')\} - 2\cdot\tilde{y}_t' D_t^{-1}\nabla_i\hat{y}_t], \tag{15}$$

where $\{\tilde{y}_t\} := \{y_t - \hat{y}_t\}$ is the so called *innovation sequence* of the Kalman filter. In a correctly specified model, this series is (Gaussian) white noise with mean zero and covariance matrix D_t (as defined in (4b)). Exploiting the moments of the innovation sequence, one can deduce the $(i,j)^{\text{th}}$ element of the information matrix as [22,23]:

$$[\mathbf{I}_N(\theta)]_{ij} = \frac{1}{2}\sum_{t=1}^{N}[\text{trace}\{D_t^{-1}\nabla_i D_t D_T^{-1}\nabla_j D_t\} + 2\cdot E(\nabla_i\hat{y}_t D_t^{-1}\nabla_j\hat{y}_t)]. \tag{16}$$

Neglecting the expectations operator in (16), one derives an estimator for the information matrix, which only depends on first-order derivatives of the moments $\hat{y}_t$ and D_t. These derivatives are taken numerically during two filtering runs in a small neighbourhood of the current iteration solution $\hat{\theta}_N^{(i)}$. The scoring method is a modified Newton method (the modification consists in substituting the Hessian matrix by its estimated expectation). Near the likelihood maximum, scoring has quadratic convergence properties, but far off the maximum this method may generate misleading search vectors due to a bad approximation of the Hessian matrix. As an alternative one may use the EM-method, which converges only linearly near the likelihood maximum [24], but which—as experience shows—also generates satisfying increases in likelihood far off the likelihood maximum. In addition the computational burden of the EM-method is far less than that of scoring. In the framework of model (1), the EM-method amounts to the solution of a standard least squares regression problem.

EM-METHOD

The general situation where the EM-method is applicable is the following: there is a joint distribution $f_{X,Y}(x,y;\theta)$ of "observables Y" and "latent variables X". We are looking for an ML-estimate of θ based upon the likelihood derived from the marginal distribution $f_Y(y;\theta)$. The EM-method starts from a decomposition of the log-likelihood into two auxiliary functions. The definition of conditional densities implies:

$$L_N(\theta) := \log f_Y(y;\theta) = \log f_{X,Y}(x,y;\theta) - \log f_{X|Y}(x \mid y;\theta).$$

If one takes expectations on both sides with respect to the distribution $f_{X|Y}(x \mid y;\bar{\theta})$, one obtains:

$$\log f_Y(y;\theta) = E[\log f_{X,Y}(x,y;\theta) \mid y,\bar{\theta}] - E[\log f_{X|Y}(x \mid y;\theta) \mid y,\bar{\theta}], \tag{17a}$$

or briefly in obvious notation:

$$\log f_Y(y;\theta) = A_{X,Y}(\theta,\bar{\theta}) - A_{X|Y}(\theta \mid \bar{\theta}). \tag{17b}$$

Starting from an iteration solution $\theta^{(i)}$, the next solution $\theta^{(i+1)}$ is constructed in two steps:

(1) *expectation step*: form the auxiliary function $A_{X,Y}(\theta,\theta^{(i)})$ (18a)

(2) *maximization step*: find $\theta^{(i+1)}$ such that for all $\theta \in \Theta$ (parameter space):

$$A_{X,Y}(\theta^{(i+1)},\theta^{(i)}) \geq A_{X,Y}(\theta,\theta^{(i)}). \tag{18b}$$

For all iteration sequences $\{\theta^{(i)}\}$ constructed according to (18), the corresponding likelihood sequence $\{L(\theta^{(i)})\}$ is nonincreasing. It will converge to a stationary point of the likelihood function under suitable continuity and differentiability conditions on $A_{X,Y}(\theta_1,\theta_2)$ (see [24,25]). The EM-method does not guarantee convergence to the global likelihood maximum. In addition one cannot exclude convergence to a saddle point; the situation has to be checked numerically by an analysis of the Hessian. The great advantage of the EM-method lies in the fact that usually the maximization problem of the auxiliary function $A_{X,Y}(\theta_1,\theta_2)$ is far simpler than direct maximization of $L(\theta)$. This is also the case in state-space models [1,26]. In this framework, the latent variables X may be identified with the state vectors $X := (X_0, X_1, \ldots, X_N)$, and the observables with the output vectors $Y := (Y_1, \ldots, Y_N)$. The joint distribution of X and Y is determined by the stochastic specification (3) as:

$$\begin{aligned}\log f_{X,Y}(x,y;\theta) &= \log f_{X(N),Y(N)}(x(N),y(N);\theta)\\ &= \log f_{Y(N)|X(N)}(y(N) \mid x(N);\theta) + \log f_{X(N)}(x(N);\theta).\end{aligned}$$

Exploiting the special Markovian structure of this econometric state-space model, we have:

$$\begin{aligned}\log f_{X,Y}(x,y;\theta) &= \sum_{t=1}^{N}[\log f(y_t \mid x_t;\theta) + \log f(x_t \mid x_{t-1};\theta)] + \log f(x_0;\theta)\\ &= \sum_{t=1}^{N}[\log \mathcal{N}(y_t : U_{1t}'x_t + U_{2t}'\gamma; R) + \log \mathcal{N}(x_t : \Phi x_{t-1} + U_{3t}'\gamma; Q)]\\ &\quad + \log \mathcal{N}(x_0 : \mu_0; \Sigma_0).\end{aligned} \tag{19}$$

Expectation Step

The auxiliary function $A_{X,Y}(\theta,\theta^{(i)})$ is the expected value of (19) with respect to the conditional distribution $f_{X(N)|Y(N)}(x(N) \mid y(N);\theta^{(i)})$ given the last iteration solution $\theta^{(i)}$. It will be convenient to decompose the function $A_{X,Y}$ into the partial sums:

$$A_{X,Y}(\theta,\theta^{(i)}) = E[\log f_{X,Y}(x,y;\theta) \mid y;\theta^{(i)}] = \sum_{j=1}^{3} A_j(\theta_j;\theta^{(i)}),$$

where

$$\begin{aligned}A_1(\theta_1,\theta^{(i)}) &= \sum_{t=1}^{N} \mathop{\mathbf{E}}_{X_t|Y(N)}[\log \mathcal{N}(y_t : U_{1t}'x_t + U_{2t}'\beta; R) \mid y(N);\theta^{(i)}]\\ &= \text{const.} - \frac{N}{2}\log|R| - \frac{1}{2}\sum_{t=1}^{N}\text{trace}\{R^{-1}E(v_tv_t' \mid y(N);\theta^{(i)})\},\end{aligned} \tag{20a}$$

$$\begin{aligned}A_2(\theta_2,\theta^{(i)}) &= \sum_{t=1}^{N} \mathop{\mathbf{E}}_{X_t,X(t-1)|Y(N)}[\log \mathcal{N}(x_t : \Phi x_{t-1} + U_{3t}'\gamma; Q) \mid y(N);\theta^{(i)}]\\ &= \text{const.} - \frac{N}{2}\log|Q| - \frac{1}{2}\sum_{t-1}^{N}\text{trace}\{R^{-1}E(w_tw_t' \mid y(N);\theta^{(i)})\},\end{aligned} \tag{20b}$$

$$\begin{aligned}A_3(\theta_3,\theta^{(i)}) &= \mathop{\mathbf{E}}_{X_0|Y(N)}[\log \mathcal{N}(x_0 : \mu_0; \Sigma_0 \mid y(N);\theta^{(i)}]\\ &= \text{const.} - \frac{1}{2}\log|\Sigma_0| - \frac{1}{2}\text{trace}\{\Sigma_0^{-1}E(w_0w_0' \mid y(N);\theta^{(i)})\}.\end{aligned} \tag{20c}$$

The cross-moments of $\{v_t\}$ and $\{w_t\}$ in (20) can be deduced from the Kalman filter recursions (4) using model specification $\theta^{(i)}$. For this purpose, the following decompositions of the noise vectors turn out to be convenient:

$$v_t = y_t - U_{1t}'x_t - U_{2t}'\beta = \hat{v}_{t|N} - U_{1t}'\tilde{x}_{t|N}, \qquad t = 1,\ldots \tag{21a}$$

$$w_t = x_t - \Phi x_{t-1} - U_{3t}'\gamma = \hat{w}_{t|N} + [\tilde{x}_{t|N} - \Phi\tilde{x}_{t-1|N}], \quad t = 1,\ldots \tag{21b}$$

$$w_0 = x_0 - \mu_0 = \hat{v}_{0|N} + \tilde{x}_{0|N}, \tag{21c}$$

where $\hat{v}_{t|N}$ and $\hat{w}_{t|N}$ are the smoothed system errors given specification $\theta^{(i)}$, i.e.,

$$\hat{v}_{t|N} := y_t - U_{1t}'\hat{x}_{t|N} - U_{2t}'\beta \qquad t = 1, 2, \ldots \tag{22a}$$

$$\hat{w}_{0|N} := \hat{x}_{0|N} - \mu_0; \quad \hat{w}_{t|N} := \hat{x}_{t|N} - \Phi\hat{x}_{t-1|N} - U_{3t}'\gamma \quad t = 1, 2, \ldots \tag{22b}$$

The prediction errors $\tilde{x}_{t|N} := x_t - \hat{x}_{t|N}$, where $\hat{x}_{t|N} = E(x_t \mid y(N); \theta^{(i)})$, have zero expectation given $y(N)$ and $\theta^{(i)}$. Since $\hat{w}_{t|N}$ and $\hat{v}_{t|N}$ are linear functions in $y(N)$, we have:

$$\begin{aligned} E(\hat{v}_{t|N}\tilde{x}_{s|N}' \mid y(N); \theta^{(i)}) &= 0 \\ E(\hat{w}_{t|N}\tilde{x}_{s|N}' \mid y(N); \theta^{(i)}) &= 0 \end{aligned} \qquad \text{for arbitrary } s, t = 0, 1, \ldots \tag{23}$$

Observing these orthogonality conditions in (21) we can complete the expectation step as

$$E(v_t v_t' \mid y(N); \theta^{(i)}) = \hat{v}_{t|N}\hat{v}_{t|N}' + U_{1t}'\Sigma_{t|N}U_{1t}, \tag{24a}$$

$$\begin{aligned} E(w_t w_t' \mid y(N); \theta^{(i)}) = {} & \hat{w}_{t|N}\hat{w}_{t|N} + \Sigma_{t|N}' - \Phi\Sigma_{t-1,t|N} \\ & - \Sigma_{t-1,t|N}'\Phi' + \Phi\Sigma_{t-1|N}\Phi', \end{aligned} \tag{24b}$$

$$E(w_0 w_0' \mid y(N); \theta^{(i)}) = \hat{w}_{0|N}\hat{w}_{0|N}' + \Sigma_{0|N}. \tag{24c}$$

Maximization Step

Obviously we can maximize (20) by maximizing each component A_j separately. Each maximization corresponds to the well-known regression problem of ***seemingly unrelated equation systems*** [27,28]. All critical points may be found by differentiation, but their actual computation usually involves the (iterative) solution of nonlinear systems. Using the parameterization (2c–d) we compute the derivatives with respect to β, γ, φ as:

$$\frac{\partial A_1}{\partial \beta} = \left[\frac{\partial A_1}{\partial \mathrm{vec}(B')'} \cdot \frac{\partial \mathrm{vec}(B')}{\partial \beta'}\right]' = C_2' \mathrm{vec}\left(\frac{\partial A_1}{\partial B'}\right), \tag{25a}$$

$$\frac{\partial A_2}{\partial \gamma} = \left[\frac{\partial A_2}{\partial \mathrm{vec}(\Gamma')'} \cdot \frac{\partial vec(\Gamma')}{\partial \gamma'}\right]' = C_3' \mathrm{vec}\left(\frac{\partial A_2}{\partial \Gamma'}\right), \tag{25b}$$

$$\frac{\partial A_2}{\partial \varphi} = \left[\frac{\partial A_2}{\partial \mathrm{vec}(\Phi')'} \cdot \frac{\partial \mathrm{vec}(\Phi')}{\partial \varphi'}\right]' = C_4' \mathrm{vec}\left(\frac{\partial A_2}{\partial \Phi'}\right). \tag{25c}$$

Observing definition (22) and result (24) we have:

$$\frac{\partial A_1}{\partial B'} = \sum_{t=1}^{N} \left[u_t y_t' - u_t \hat{x}_{t|N}' U_{1t} - u_t u_t' B'\right] R^{-1}, \tag{26a}$$

$$\frac{\partial A_2}{\partial \Phi'} = \sum_{t=1}^{N} \left[M_{t-1,t|N} - \hat{x}_{t-1|N}\, u_t'\Gamma - M_{t-1|N}\Phi'\right] Q^{-1}, \tag{26b}$$

$$\frac{\partial A_2}{\partial \Gamma'} = \sum_{t=1}^{N} \left[u_t \hat{x}_{t|N}' - u_t \hat{x}_{t-1|N}'\Phi' - u_t u_t'\Gamma'\right] Q^{-1}, \tag{26c}$$

$$\frac{\partial A_1}{\partial R} = -\frac{N}{2} + \frac{1}{2}\sum_{t=1}^{N} R^{-1} E(v_t v_t' \mid y(N); \theta^{(i)}) R^{-1}, \tag{26d}$$

$$\frac{\partial A_2}{\partial Q} = -\frac{N}{2} + \frac{1}{2}\sum_{t=1}^{N} Q^{-1} E(w_t w_t' \mid y(N); \theta^{(i)}) Q^{-1}, \tag{26e}$$

$$\frac{\partial A_3}{\partial \mu_0'} = (\hat{x}_{0|N} - \mu_0)'\Sigma_0^{-1}, \tag{26f}$$

$$\frac{\partial A_3}{\partial \Sigma_0} = -\frac{1}{2}\Sigma_0^{-1}\left[M_{0|N} - \hat{x}_{0|N}\mu_0' - \mu_0\hat{x}_{0|N}' + \mu_0\mu_0'\right]\Sigma_0^{-1}, \tag{26g}$$

where $M_{t-1|N}$ and $M_{t-1,t|N}$ are the smoothed cross-moments of the states, which are according to (9)

$$M_{t-1|N} := E(x_{t-1}x'_{t-1} \mid y(N);\theta^{(i)}) = \Sigma_{t-1|N} + \hat{x}_{t-1|N}\,\hat{x}'_{t-1|N}, \tag{27a}$$

$$M_{t-1,t|N} := E(x_{t-1}x'_t \mid y(N);\theta^{(i)}) = \Sigma_{t-1|t-1}\Phi'\Sigma^{-1}_{t|t-1}\Sigma_{t|N} + \hat{x}_{t-1|N}\,\hat{x}'_{t|N}. \tag{27b}$$

Setting all derivatives to zero, we arrive at the final (nonlinear) set of equations determining the EM-solution $(\hat{\beta},\hat{\gamma},\hat{\varphi},\hat{R},\hat{Q},\hat{\mu}_0,\hat{\Sigma}_0)$ given model specification $\theta^{(i)}$ (assumed in the computation of all smoothed moments used below):

$$C'_2(\hat{R}^{-1}\otimes I_n)\,\mathrm{vec}\left(\sum_{t=1}^{N} u_t(y_t - U'_{1t}\hat{x}_{t|N})'\right) = \left[C'_2\left(\hat{R}^{-1}\otimes\sum_{t=1}^{N}u_t u'_t\right)C_2\right]\cdot\hat{\beta}, \tag{28a}$$

$$\begin{bmatrix} C'_3(\hat{Q}^{-1}\otimes I_n)\mathrm{vec}\left(\sum_{t=1}^{N} u_t\hat{x}'_{t|N}\right) \\ C'_4(\hat{Q}^{-1}\otimes I_k)\mathrm{vec}\left(\sum_{t=1}^{N} M_{t-1,t|N}\right)\end{bmatrix} = \begin{bmatrix} C'_3\left(\hat{Q}^{-1}\otimes\sum_{t=1}^{N}u_t u'_t\right)C_3 & C'_3\left(\hat{Q}^{-1}\otimes\sum_{t=1}^{N}u_t\hat{x}'_{t-1|N}\right)C_4 \\ C'_4\left(\hat{Q}^{-1}\otimes\sum_{t=1}^{N}\hat{x}_{t-1|N}u'_t\right)C_3 & C'_4\left(\hat{Q}^{-1}\otimes\sum_{t=1}^{N}M_{t-1|N}\right)C_4\end{bmatrix}\cdot\begin{bmatrix}\hat{\gamma}\\ \hat{\varphi}\end{bmatrix}, \tag{28b}$$

$$\hat{\mu}_0 = \hat{x}_{0|N}, \qquad \hat{\Sigma}_0 = \Sigma_{0|N}, \tag{28c}$$

$$\hat{R} = \frac{1}{N}\sum_{t=1}^{N} E(v_t v'_t \mid y(N),\theta^{(i)}), \tag{29a}$$

$$\hat{Q} = \frac{1}{N}\sum_{t=1}^{N} E(w_t w'_t \mid y(N),\theta^{(i)}). \tag{29b}$$

The solution for μ_0,Σ_0 is trivial, since it coincides with the smoothing solution. For the remaining parameters we can exploit the usual estimation techniques available for systems of seemingly unrelated regression equations [27], i.e., two-step or iterated AITKEN estimators. The two-step AITKEN estimator consists in updating φ, γ, β given the last iteration solution for R and Q and then substituting these updates into (28) in order to generate a new solution of R and Q, i.e.,

$$\hat{\beta}^{(i+1)} = \left[C'_2\left([\hat{R}^{(i)}]^{-1}\otimes\sum_{t=1}^{N}u_t u'_t\right)C_2\right]^{-1}\left[C'_2\left([\hat{R}^{(i)}]^{-1}\otimes I_n\right)\mathrm{vec}\left(\sum_{t=1}^{N}u_t\left(y_t - U'_{1t}\hat{x}^{(i)}_{t|N}\right)\right)\right], \tag{30a}$$

$$\begin{bmatrix}\hat{\gamma}^{(i+1)}\\ \hat{\varphi}^{(i+1)}\end{bmatrix} = \begin{bmatrix} C'_3\left([\hat{Q}^{(i)}]^{-1}\otimes\sum_{t=1}^{N}u_t u'_t\right)C_3 & C'_3\left([\hat{Q}^{(i)}]^{-1}\otimes\sum_{t=1}^{N}u_t\hat{x}^{(i)\prime}_{t-1|N}\right)C_4 \\ C'_4\left([\hat{Q}^{(i)}]^{-1}\otimes\sum_{t=1}^{N}\hat{x}^{(i)}_{t-1|N}u'_t\right)C_3 & C'_4\left([\hat{Q}^{(i)}]^{-1}\otimes\sum_{t=1}^{N}M^{(i)}_{t-1|N}\right)C_4\end{bmatrix}^{-1} \times \begin{bmatrix} C'_3([\hat{Q}^{(i)}]^{-1}\otimes I_n)\,\mathrm{vec}\left(\sum_{t=1}^{N}u_t\hat{x}^{(i)\prime}_{t|N}\right) \\ C'_4([\hat{Q}^{(i)}]^{-1}\otimes I_k)\mathrm{vec}\left(\sum_{t=1}^{N}M^{(i)}_{t-1,t|N}\right)\end{bmatrix}. \tag{30b}$$

$$\hat{R}^{(i+1)} = \frac{1}{N}\sum_{t=1}^{N}\left(\hat{v}_{t|N}^{(i)}\hat{v}_{t|N}^{(i)\prime} + U_{1t}'\sum\nolimits_{t|N}^{(i)}U_{1t}\right), \tag{31a}$$

$$\hat{Q}^{(i+1)} = \frac{1}{N}\sum_{t=1}^{N}\left(\hat{w}_{t|N}^{(i)}w_{t|N}^{(i)\prime} + \mathcal{W}_{t|N}^{(i)}\right), \tag{31b}$$

where

$$\hat{v}_{t|N}^{(i)} := y_t - U_{1t}'\hat{X}_{t|N}^{(i)} - U_{2t}'\hat{\beta}^{(i+1)}, \tag{32a}$$

$$\hat{w}_{t|N}^{(i)} := \hat{x}_{t|N}^{(i)} - \hat{\Phi}^{(i+1)}\hat{x}_{t-1|N}^{(i)} - U_{3t}'\hat{\gamma}^{(i+1)}, \tag{32b}$$

$$\mathcal{W}_{t|N}^{(i)} := \Sigma_{t|N}^{(i)} + \hat{\Phi}^{(i+1)}\Sigma_{t-1|N}^{(i)}\hat{\Phi}^{(i+1)\prime} - \hat{\Phi}^{(i+1)}\Sigma_{t-1,t|N}^{(i)} - \Sigma_{t-1,t|N}^{(i)\prime}\hat{\Phi}^{(i+1)\prime}. \tag{32c}$$

Note that $\mathcal{W}_{t|N}^{(i)}$, as defined in (32c), should be positive definite. It is a quadratic form in the smoothed covariance matrix of $\begin{pmatrix} x_{t-1} \\ x_t \end{pmatrix}$. We have:

$$\begin{aligned}\mathcal{W}_{t|N}^{(i)} &:= \mathrm{COV}(\tilde{x}_{t|N} - \Phi\tilde{x}_{t-1|N} \mid y(N);\theta^{(i)}) \\ &= (\hat{\Phi}^{(i+1)} - I)\cdot\begin{pmatrix}\Sigma_{t-1|N}^{(i)} & \Sigma_{t-1,t|N}^{(i)} \\ \Sigma_{t-1,t|N}^{(i)\prime} & \Sigma_{t|N}^{(i)}\end{pmatrix}\cdot\begin{pmatrix}\hat{\Phi}^{(i+1)\prime} \\ -I\end{pmatrix}.\end{aligned} \tag{33}$$

The iterated AITKEN estimator is computed iterating between (30) and (31). The convergence of these (sub)iterations are guaranteed if the matrix series $\{\hat{R}, \hat{Q}\}$, obtained during these subiterations, can be uniformly bounded from above and below by positive definite matrices [29]. The EM-iterations will generate increases in the likelihood in each step. If there are no ridges in the likelihood, they will converge to a stationary point of the likelihood [25]. Since convergence to a saddle point or to the global maximum is not guaranteed a final step of scoring is required to check the curvature of the likelihood at that point. The Hessian or an approximation of it is also needed for generating confidence intervals for the complete set of hyperparameters θ.

ADAPTIVE EM ESTIMATION

There is an obvious way to turn the EM-algorithm into an on-line estimation technique for all hyperparameters. This is simply done by approximating the *smoothing* solutions $\hat{x}_{t|N}$, $\Sigma_{t|N}$ by the corresponding *filtering* solutions available at time t. Substituting t for N in all subindices of (30) and (31), denoting the information set being conditioned upon, we arrive at a series of equation systems which can be built up recursively during one filtering run. Making the relevant substitutions we have for $N = \bar{N}, \bar{N}+1, \ldots$ (where $\bar{N}$ is set in such a way, that invertibility in the formulas below is assured):

$$\hat{\beta}^{(N+1)} = \left[C_2'\left([\hat{R}^{(N)}]^{-1}\otimes\sum_{t=1}^{N}u_tu_t'\right)C_2\right]^{-1}\left[C_2'\left([\hat{R}^{(N)}]^{-1}\otimes I_n\right)\mathrm{vec}\left(\sum_{t=1}^{N}u_t(y_t - U_{1t}'\hat{x}_{t|t}^{(t)})\right)\right], \tag{34a}$$

$$\begin{bmatrix}\hat{\gamma}^{(N+1)} \\ \hat{\varphi}^{(N+1)}\end{bmatrix} = \begin{bmatrix} C_3'\left([\hat{Q}^{(N)}]^{-1}\otimes\sum_{t=1}^{N}u_tu_t'\right)C_3 & C_3'\left([\hat{Q}^{(N)}]^{-1}\otimes\sum_{t=1}^{N}u_t\hat{x}_{t-1|t}^{(t)\prime}\right)C_4 \\ C_4'\left([\hat{Q}^{(N)}]^{-1}\otimes\sum_{t=1}^{N}\hat{x}_{t-1|t}^{(t)}u_t'\right)C_3 & C_4'\left([\hat{Q}^{(N)}]^{-1}\otimes\sum_{t=1}^{N}M_{t-1|t}^{(t)}\right)C_4\end{bmatrix}^{-1} \begin{bmatrix} C_3'\left([\hat{Q}^{(N)}]^{-1}\otimes I_n\right)\mathrm{vec}\left(\sum_{t=1}^{N}u_t\hat{x}_{t|t}^{(t)\prime}\right) \\ C_4'\left([\hat{Q}^{(N)}]^{-1}\otimes I_k\right)\mathrm{vec}\left(\sum_{t=1}^{N}M_{t-1,t|t}^{(t)}\right)\end{bmatrix}, \tag{34b}$$

$$\hat{R}^{(N+1)} = \frac{1}{N}\sum_{t=1}^{N}\left(\hat{v}_{t|t}^{(t)}\,\hat{v}_{t|t}^{(t)\prime} + U_{1t}'\Sigma_{t|t}^{(t)}U_{1t}\right), \tag{35a}$$

$$\hat{Q}^{(N+1)} = \frac{1}{N}\sum_{t=1}^{N}\left(\hat{w}_{t|t}^{(t)}w_{t|t}^{(t)\prime} + \mathcal{W}_{t|t}^{(t)}\right), \tag{35b}$$

where

$$\hat{v}_{t|t}^{(t)} := y_t - U_{1t}'\hat{x}_{t|t}^{(t)} - U_{2t}'\hat{\beta}^{(t+1)}, \tag{36a}$$

$$\hat{w}_{t|t}^{(t)} := \hat{x}_{t|t}^{(t)} - \hat{\Phi}^{(t+1)}\hat{x}_{t-1|t}^{(t)} - U_{3t}'\hat{\gamma}^{(t+1)}, \tag{36b}$$

$$\mathcal{W}_{t|t}^{(t)} := \Sigma_{t|t}^{(t)} + \hat{\Phi}^{(t+1)}\Sigma_{t-1|t}^{(t)}\hat{\Phi}^{(t+1)\prime} - \hat{\Phi}^{(t+1)}\Sigma_{t-1,t|t}^{(t)} - \Sigma_{t-1,t|t}^{(t)\prime}\hat{\Phi}^{(t+1)\prime}. \tag{36c}$$

All estimators can be computed in a fully recursive fashion for $N = \bar{N}, \bar{N}+1, \ldots$ using the Kalman filter recursions (4) as well as those for the so-called ***one-step back smoother***, which is just a special version of (7) (substituting t for N) [9, pp. 187–190]:

$$\hat{x}_{t-1|t} = \hat{x}_{t-1|t-1} + A_{t-1}(\hat{x}_{t|t} - \hat{x}_{t|t-1}), \tag{37a}$$

$$\Sigma_{t-1|t} = \Sigma_{t-1|t-1} + A_{t-1}(\Sigma_{t|t} - \Sigma_{t|t-1})A_{t-1}', \tag{37b}$$

where

$$A_{t-1} := \Sigma_{t-1|t-1}\Phi'\Sigma_{t|t-1}^{-1}.$$

The superscript (t) for the state moments in (35) and (36) indicates that the most recent model specification $\theta^{(t)}$ is used for the recursions (4) and (37). Approximating the usual EM-method in this way substantially reduces the storage requirements for intermediate results.

Some comments on the character of the variance estimators follow. Note that the decomposition (21) is also valid for $N = t$, whence we have:

$$E(v_t v_t' \mid y(t)) = \hat{v}_{t|t}\,\hat{v}_{t|t}' + U_{1t}'\Sigma_{t|t}U_{1t}, \tag{38a}$$

$$E(w_t w_t' \mid y(t)) = \hat{w}_{t|t}\,\hat{w}_{t|t}' + \mathcal{W}_{t|t}, \tag{38b}$$

where

$$\mathcal{W}_{t|t} = \Sigma_{t|t}' - \Phi\Sigma_{t-1,t|t} - \Sigma_{t-1,t|t}'\Phi' + \Phi\Sigma_{t-1|t}\Phi'. \tag{38c}$$

Taking expected values on both sides of (36), we get in a correctly specified state space:

$$R = E\left(\frac{1}{N}\sum_{t=1}^{N}\hat{v}_{t|t}\hat{v}_{t|t}'\right) + \frac{1}{N}\sum_{t=1}^{N}U_{1t}'\Sigma_{t|t}U_{1t}, \tag{39a}$$

$$Q = E\left(\frac{1}{N}\sum_{t=1}^{N}\hat{w}_{t|t}\hat{w}_{t|t}'\right) + \frac{1}{N}\sum_{t=1}^{N}\mathcal{W}_{t|t}. \tag{39b}$$

Hence, the variance estimator (35) may be interpreted as a special "method-of-moments estimator", where $E(\frac{1}{N}\sum_{t=1}^{N}\hat{v}_{t|t}\hat{v}_{t|t}')$ and $E(\frac{1}{N}\sum_{t=1}^{N}\hat{w}_{t|t}\hat{w}_{t|t}')$ are substituted by the corresponding observed values and where all moments have been calculated as the filtering solution based on some prior (or recursively updated as in (34)) values for the state space hyperparameters. This estimator is very closely related to an adaptive filter proposed by [2], whose approach amounts to the solution of a slightly modified version of (39), namely:

$$\frac{1}{N}\sum_{t=1}^{N}\left(I - U_{1t}'\Sigma_{t|t}U_{1t}R_0^{-1}\right)R = \frac{1}{N}\sum_{t=1}^{N}\hat{v}_{t|t}\,\hat{v}_{t|t}', \tag{40a}$$

$$\frac{1}{N}\sum_{t=1}^{N}(I - \mathcal{W}_{t|t}Q_0^{-1})Q = \frac{1}{N}\sum_{t=1}^{N}\hat{w}_{t|t}\,\hat{w}_{t|t}', \tag{40b}$$

where Q_0 and R_0 are prior values for Q and R, "not too different" from the true specification. Louv actually only equates the diagonal elements of the matrix equation (40) and computes the filtering solutions for a fixed prior specification, where Φ is known and β, γ are zero. This procedure may be interpreted as an approximation to Rao's ***minimum norm quadratic unbiased estimators*** (MINQUE) for the variance components of R and Q in the stochastic setup of state space model (1), as we will now show.

RELATIONSHIP BETWEEN EM- AND MINQUE-ESTIMATORS OF (CO)VARIANCE COMPONENTS

We want to concentrate on the estimation of the variance components of R and Q, therefore we consider a simplified version of (1), where Φ is known and γ, β are set to zero. The resulting measurement and transition equations can be collected into a regression model with random coefficients of the following form:

$$\begin{bmatrix} \mu_0 \\ 0 \\ y_1 \\ 0 \\ y_2 \\ \vdots \\ 0 \\ y_N \end{bmatrix} = \begin{bmatrix} -I & I & 0 & 0 & \dots & 0 & 0 \\ 0 & -\Phi & I & 0 & & 0 & 0 \\ 0 & 0 & U_1' & 0 & \dots & 0 & 0 \\ 0 & 0 & -\Phi & I & & 0 & 0 \\ 0 & 0 & 0 & U_2' & & 0 & 0 \\ \vdots & & \vdots & & \ddots & & \vdots \\ 0 & 0 & 0 & 0 & & -\Phi & I \\ 0 & 0 & 0 & 0 & \dots & 0 & U_N' \end{bmatrix} \cdot \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_{N-1} \\ x_N \end{bmatrix} = \begin{bmatrix} -w_0 \\ -w_1 \\ v_1 \\ -w_2 \\ v_2 \\ \vdots \\ -w_N \\ v_N \end{bmatrix}, \tag{41a}$$

or briefly in obvious notation ($M = N \cdot (m+k) + k$, $K = N \cdot k$):

$$\underset{(M\times 1)}{y} = \underset{(M\times K)}{H'} \cdot \underset{(K\times 1)}{x} + \underset{(M\times 1)}{\epsilon}, \quad \text{where} \quad E(\epsilon) = 0, \quad E(\epsilon\epsilon') = \Sigma, \tag{41b}$$

Σ being blockdiagonal of the form:

$$\Sigma = \operatorname{diag}(\Sigma_0, Q, R, Q, R, \dots, Q, R). \tag{41c}$$

The *minimum variance linear unbiased estimator* (MVLUE) for the random variable x is characterized by the following conditions:

$$\hat{x} = Ay \qquad \text{linearity in } y, \tag{42a}$$

$$E(\hat{x} - x) = 0 \qquad \text{unbiasedness}, \tag{42b}$$

$$\operatorname{VAR}(\lambda'(\hat{x} - x)) \quad \text{is minimum among all linear unbiased estimators of } x \text{ for any given } \lambda \in \mathbf{R}^K, \tag{42c}$$

and is uniquely given by the generalized AITKEN estimator [30,31]:

$$\hat{x} = (H\Sigma^{-1}H')^{-1}(H\Sigma^{-1}y). \tag{43}$$

This estimator $\hat{x}$ may equivalently be characterized as the minimum least squares estimator of x. It coincides with the smoothing solution $\hat{x} = (\hat{x}_{0|N}', \dots, \hat{x}_{N|N}')'$ given Σ [32]. Given a prior estimate S of Σ, we compute an approximate MVLUE of x as $\hat{x}_S$ (the smoothing solution using a suboptimal filter corresponding to S):

$$\hat{x}_S := (HS^{-1}H')^{-1}(HS^{-1}y). \tag{44}$$

The respective least squares error vector is $\hat{\epsilon}_S$:

$$\hat{\epsilon}_S := y - H'\hat{x}_S = (I - K_S)\epsilon, \tag{45}$$

where

$$K_S := H'(HS^{-1}H')^{-1}HS^{-1}. \tag{46}$$

The vector $\hat{\epsilon}_S$ contains the smoothed errors $\{\hat{w}_{t|N}, \hat{v}_{t|N}\}$ using a suboptimal filter. The MINQUE-estimator for the variance components of Σ (see [33] and [34]) is essentially a method of moments estimator constructed in the following way. Compute the population cross moments of $\hat{\epsilon}_S$ as

$$E(\hat{\epsilon}_S\,\hat{\epsilon}_S') = (I - K_S)\Sigma(I - K_S)', \tag{47}$$

replace the expected value in (47) by the corresponding observed value and equate those elements in the resulting matrix equation, which correspond to the nonzero elements of Σ, then collect terms and solve for the variance components of Σ. This equation system can be simplified considerably, if we assume that $\Sigma \approx S$. Observing (45), we have approximately:

$$E(\hat{\epsilon}_S \hat{\epsilon}_S') = \Sigma - H'(HS^{-1}H')^{-1}H. \tag{48}$$

Replacing the expected values, as usual, by the corresponding observed value, we arrive at an approximate MINQUE-equation system of the form:

$$\hat{\epsilon}_S \hat{\epsilon}_S' = \Sigma - H'(HS^{-1}H')^{-1}H. \tag{49}$$

Note that because of the special structure of H, we can collect the terms in the t^{th} $(k+m) \times (k+m)$ block on the diagonal of $H(HS^{-1}H')^{-1}H'$ (corresponding to the error terms (w_t, v_t)) and write the matrix equation resulting from (49) as:

$$\sum_{t=1}^{N} \begin{bmatrix} \hat{w}_{t|N}\hat{w}_{t|N}' & \hat{w}_{t|N}\hat{v}_{t|N} \\ \hat{v}_{t|N}\hat{w}_{t|N}' & \hat{v}_{t|N}\hat{v}_{t|N}' \end{bmatrix} = \sum_{t=1}^{N} \begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix} + \sum_{t=1}^{N} \begin{bmatrix} \Phi\Sigma_{t-1|N}\Phi' - \Sigma_{t-1,t|N}'\Phi' - \Phi\Sigma_{t-1,t|N} + \Sigma_{t|N} & (\Sigma_{t|N} - \Phi\Sigma_{t-1,t|N})U_t \\ U_t'(\Sigma_{t|N} - \Phi\Sigma_{t-1,t|N})' & U_t'\Sigma_{t|N}U_t \end{bmatrix}. \tag{50}$$

The *EM-method* of variance component estimation consists in solving equation (49) as:

$$\hat{\epsilon}_S \hat{\epsilon}_S' + H'(HS^{-1}H')^{-1}H = \hat{\Sigma}. \tag{51}$$

The *AUE-method* of variance component estimation [the name is due to an approach of Horn *et al.* [35], who called their approximate MINQUE an *almost unbiased estimator*] consists in solving equation (49) [2, p. 401] as

$$\hat{\epsilon}_S \hat{\epsilon}_S' = (I - H'(HS^{-1}H')^{-1}HS^{-1})\hat{\Sigma} = (I - K_S)\hat{\Sigma} \tag{52}$$

Observing (50), we see that collecting terms in (51) and (52), equating only the matrix elements corresponding to the blockdiagonal of Σ and finally approximating smoothing by filtering solutions amounts to the estimation procedures based on (39) and (40) as discussed in the last section.

A SIMULATION STUDY

In the following study, we analyze the behaviour of the methods described above within a special version of (1) known as the "convergent parameter model" or "return to normality model" in the econometrics literature [36]. The measurement equation is taken to be:

$$y_t = u_{11,t} \cdot \beta_{1t} + u_{12,t} \cdot \beta_{2t} + v_t, \tag{53a}$$

where β_{1t} and β_{2t} have a tendency to return to their means $\bar{\beta}_1$ and $\bar{\beta}_2$ over time, in particular:

$$(\beta_{it} - \bar{\beta}_i) = \varphi_i(\beta_{i,t-1} - \bar{\beta}_i) + w_{it}, \quad 0 < \varphi_i < 1, \quad i = 1, 2. \tag{53b}$$

For an example of this kind of model, imagine y_t to be a return on some stock S over holding period t, u_{1t} to be the return on a stock market index in period t and u_{2t} some other major economic factor influencing the return on S. The coefficient β_{1t} is called the *beta coefficient of S* in capital market theory and plays a major rule in the portfolio decision of the investor.

Define the state to be $x_{it} := \beta_{it} - \bar{\beta}_i$ for $i = 1, 2$. Then (53) may be cast into the state space:

$$y_t = (u_{11,t} \quad u_{12,t}) \cdot \begin{pmatrix} x_{1t} \\ x_{2t} \end{pmatrix} + (u_{11,t} \quad u_{12,t}) \cdot \begin{pmatrix} \bar{\beta}_1 \\ \bar{\beta}_2 \end{pmatrix} + v_t, \tag{54a}$$

$$\begin{pmatrix} x_{1t} \\ x_{2t} \end{pmatrix} = \begin{pmatrix} \varphi_1 & 0 \\ 0 & \varphi_2 \end{pmatrix} \cdot \begin{pmatrix} x_{1,t-1} \\ x_{2,t-1} \end{pmatrix} + \begin{pmatrix} w_{1t} \\ w_{2t} \end{pmatrix}. \tag{54b}$$

The asymptotics of this type of model have been extensively studied in a seminal paper by [23]; under the stochastic setup given above all parameters are identified and the ML estimators are known to be consistent and asymptotically normal.

We have tried out several parameter constellation for (54). The main features of the simulation results did not change. The results presented below are based on 100 replications of a $N = 100$ observation model of (54) using the "true" parameters:

$$\bar{\beta}_1 = 10, \quad \bar{\beta}_2 = 20, \quad \varphi_1 = 0.1, \quad \varphi_2 = 0.9, \quad R = 2, \quad Q = \begin{pmatrix} 4 & 0 \\ 0 & 8 \end{pmatrix}.$$

Consistent estimates of $\bar{\beta}_i$ are available by OLS; these (inefficient) estimates were used as starting values. All other parameters have been set to unity. The regressors $u_{1i,t}$ are standard normal variates, the other error terms were constructed from standard normal variates multiplied by the respective standard deviations. We used the random number generator implemented in GAUSS 2.0—the programming language, in which all computational procedures have been written. Scoring, EM and adaptive EM have been performed on this artifical data set, in order to calculate the ML-estimators of θ, consisting of $\bar{\beta}$ and the diagonals of Φ, Q and R. Between the iteration runs, the values for μ_0 and Σ_0 were updated by a smoothing run. (Anyway, using a stable transition matrix guarantees that these starting values do not exert a high influence on the likelihood.)

The iterations have been stopped on the basis of the relative stopping rule: whenever the relative likelihood increase or the relative change in the Euclidean norm of the parameter estimate was less than one percent, the likelihood maximum was assumed to have been achieved.

Tables 1–4 present a quantitative picture of the more qualitative remarks in the text ($N = 100$ observations and 7 unknown parameters to be estimated):

(1) Scoring converges quadratically whereas the others only do linearly: the number of scoring iterations lie between 4 and 6, whereas between 15 and 35 iterations are required for the EM methods to reach the likelihood maximum. The adaptive variant of EM can be considerably faster.

(2) Scoring achieved 96% of the total likelihood increase during the first two iterations, EM and adaptive EM come very close to this performance by achieving 90% and 88%, respectively, during the first two steps. Considering the fact that one scoring iteration took ten times longer than a full EM and 17 times longer than an adaptive EM iteration, this is a strong argument in favour of an estimation start by one of the latter two methods. (The time difference grows exponentially with additional parameters to be estimated.)

(3) Scoring is more precise; this fact is reflected in a slightly higher final likelihood achieved and in more favourable error distributions of the estimators. The likelihood tended to be very flat near the optimum, and scoring seems to be better at coping with the low curvature at the optimum than the EM methods. There is also some loss in precision between the full and adaptive variant of the EM-method: the estimation error variances tend to be higher for adaptive EM throughout.

ADAPTIVE ESTIMATION BY STATE EXPANSION

Another straightforward way to construct on-line estimates for β and γ is to make them state variables, which amounts to a re-specification of (1) as:

$$y_t = H_t^{*\prime} x_t^* + v_t, \tag{55a}$$

$$x_t = \Phi_t^* x_{t-1}^* + G w_t, \tag{55b}$$

where

$$x_t^* := \begin{bmatrix} x_t \\ \beta_t \\ \gamma_t \end{bmatrix}, \quad H_t^* := \begin{bmatrix} U_{1t} \\ U_{2t} \\ 0 \end{bmatrix}, \quad G := \begin{bmatrix} I_k \\ 0 \\ 0 \end{bmatrix}, \quad \Phi_t^* := \begin{bmatrix} \Phi & 0 & U_{3t'} \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix}. \tag{56}$$

Table 1. Distribution of the numbers of iterations.

method	scoring	full EM	adaptive EM
means	4.73	16.95	14.83
std. dev.	0.764	6.61	4.88

Table 2. Distribution of the likelihood increases.

	method	scoring	full EM	adaptive EM
step 1	means	81.429	73.322	70.773
	std. dev.	22.571	21.356	21.614
step 2	means	10.295	11.024	10.448
	std. dev.	4.662	1.981	2.452
step 3	means	2.404	3.423	4.021
	std. dev.	1.147	0.624	1.206
step 4	means	0.590	1.505	2.189
	std. dev.	0.679	0.366	0.831
step 5	means	0.027	0.853	1.384
	std. dev.	0.095	0.270	0.654
final lik.	means	−170.807	−172.013	−172.556
	std. dev.	7.817	7.712	7.622

Table 3. Average stepwise likelihood increase in percent of total likelihood increase achieved.

method	scoring	full EM	adaptive EM
step 1	0.857	0.784	0.761
step 2	0.110	0.121	0.116
step 3	0.026	0.038	0.045
step 4	0.006	0.017	0.024
step 5	0.0004	0.009	0.015

Table 4. Error distribution of parameter estimates.

	method	scoring	full EM	adaptive EM
$\bar{\beta}_1$	means	−0.023	0.028	0.045
	std. dev.	0.372	0.365	0.441
$\bar{\beta}_2$	means	0.301	−0.210	−0.143
	std. dev.	2.649	3.020	3.223
φ_1	means	0.089	0.341	0.214
	std. dev.	0.355	0.221	0.328
φ_2	means	−0.042	−0.038	−0.039
	std. dev.	0.079	0.061	0.079
Q_{11}	means	−0.336	−1.981	−0.950
	std. dev.	2.046	1.131	1.540
Q_{22}	means	−0.454	−0.900	−0.955
	std.dev.	2.525	2.559	3.098
R	means	0.123	1.007	0.317
	std. dev.	1.150	1.066	1.081

The stochastics are specified as in (3) except for the initial state; the distribution of which now also includes prior means and variances for β and γ. Scoring and EM may then be applied to the extended state-space model. This technique, of course, cannot be readily applied to an

estimation of Φ and of the variance components in Q and R, since this would lead to a model which is nonlinear in the states. Extended Kalman filtering, however, is available for such a problem [37]. For an application in an econometric context, see [38].

CONCLUDING REMARKS

We mentioned, in the first section of the paper, that the statistical analysis was going to encompass also the derivation of precision measures for states and the hyperparameters. As for the latter, a final step of scoring in all components of θ yields an estimate of the information matrix and thus under asymptotic normality asymptotic confidence regions for θ. (For a survey of the asymptotic properties of ML estimators in econometric state space models, see [39].) As for the states, one might be tempted to quote the series $\{\Sigma_{t|N}\}$ running a filter using the ML-estimate of θ. This, however, would treat $\hat{\theta}$ as being a certain quantity. For a Bayesian discussion of a better risk measure, also taking account of the uncertainty due to model misspecification, see [40] which applies parallel Kalman filtering.

REFERENCES

1. M.W. Watson and R.F. Engle, Alternative algorithms for the estimation of dynamic factor, mimic and varying coefficient regression models, *J. of Econometrics* **23**, 385–400 (1983).
2. W.C. Louv, Adaptive filtering, *Technometrics* **26**, 399–409 (1984).
3. M. Aoki, *State Space Modeling of Time Series*, Springer, Berlin, (1987).
4. H. Ruskeepää, Conditionally Gaussian distributions and an application to Kalman filtering with stochastic regressors, *Communications in Statistics* **A14**, 2919–2942 (1985).
5. H. Ruskeepää, *Kalman Filtering and Prediction for Discrete-Time Models with Stochastic Regressions*, Publications of the Institute for Applied Mathematics, No 13, University of Turku, Turku, Finland, (1986).
6. A.H. Sarris, Kalman filter models: A Bayesian approach to estimation of time varying regression coefficients, *Ann. of Econ. and Soc. Measurement* **2**, 501–523 (1973).
7. P.J. Harrison and C.F. Stevens, Bayesian forecasting (with discussion), *J. of the Royal Stat. Soc.* **B38**, 205–247 (1976).
8. R.J. Meinhold and N. Singpurwalla, Understanding the Kalman filter, *The Amer. Statistican* **37**, 123–127.
9. B.D.O. Anderson and J.B. Moore, *Optimal Filtering*, Prentice-Hall, Englewood Cliffs, New Jersey, (1979).
10. P.G. Kaminski, A.E. Bryson and S.F. Schmidt, Discrete square root filtering: A survey of current techniques, *IEEE Trans. on Autom. Control* **AC–16**, 727–736 (1971).
11. M. Morf and T. Kailath, Square-root algorithms for least-squares estimation, *IEEE Trans. on Autom. Control* **AC–20**, 487–497 (1975).
12. G.J. Bierman, Sequential square root filtering and smoothing of discrete linear systems, *Automatica* **10**, 147–158 (1974).
13. G.J. Bierman, *Factorization Methods for Discrete Sequential Estimation*, Academic Press, New York, (1977).
14. D.C. Fraser and J.E. Potter, The optimum linear smoother as a combination of two optimum linear filters, *IEEE Trans. on Autom. Control* **14**, 387–390 (1969).
15. W. Schneider, *Der Kalmanfilter als Instrument zur Diagnose und Schätzung variabler Parameter in ökonometrischen Modellen*, Physica, Heidelberg, (1986).
16. C.F. Ansley and R. Kohn, A geometrical derivation of the fixed interval smoothing algorithm, *Biometrika* **69**, 486–487 (1982).
17. T.F. Cooley, B. Rosenberg and K.D. Wall, A note on the optimal smoothing for time varying coefficient problems, *Ann. of Econ. and Soc. Measurement* **6**, 453–262 (1977).
18. B. Rosenberg, Estimation of error covariance in regression with sequentially varying parameters, *Ann. of Economic and Social Measurement* **6**, 457–462 (1977).
19. C.R. Rao, *Linear Statistical Inference and Its Applications*, John Wiley and Sons, New York, (1973).
20. L.S. Lasdon, *Optimization Theory for Large Systems*, Wiley, New York, (1970).
21. J.R. Magnus and H. Neudecker, *Matrix Differential Calculus*, Wiley, New York, (1988).
22. N.K. Gupta and R.K. Mehra, Computational aspects of maximum likelihood estimation and reduction of sensitivity function calculations, *IEEE Trans. on Autom. Control* **AC-19**, 774–783 (1974).
23. A. Pagan, Some identification and estimation results for regression models with stochastically varying coefficients, *J. of Econometrics* **13**, 341–363 (1980).
24. A.P. Dempster, N.M. Laird and D.B. Rubin, Maximum-likelihood from incomplete data via the EM-algorithm, *J. of the Royal Stat. Soc.* **B39**, 1–38 (1977).
25. C.F.J. Wu, On the convergence properties of the EM-algorithm, *Ann. of Statistics* **11**, 95–103 (1983).
26. R.H. Shumway and D.S. Stoffer, An approach to time series smoothing and forecasting using the EM-algorithm, *J. of Time Series Analysis* **3**, 253–264 (1982).
27. A. Zellner, An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias, *J. Amer. Stat. Assoc.* **57**, 348–368 (1962).

28. A. Zellner, Estimators for seemingly unrelated regression equations: Some exact finite sample results, *J. Amer. Stat. Assoc.* **58**, 977–992 (1963).
29. W. Oberhofer and J. Kmenta, A general procedure for obtaining maximum likelihood estimates via generalized regression models, *Econometrica* **42**, 579–590 (1974).
30. C.R. Rao, The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves, *Biometrika* **52**, 447-458 (1965).
31. D. Pfeffermann, On extensions of the Gauss-Markov Theorem to the case of stochastic regression coefficients, *J. Royal Stat. Society* **B46**, 139–148 (1984).
32. D.B. Duncan and S.D. Horn, Linear recursive estimation from the viewpoint of regression analysis, *J. of the Amer. Stat. Assoc.* **67**, 815–821 (1972).
33. C.R. Rao, Estimation of heteroskedastic variances in linear models, *J. of the Amer. Stat. Assoc.* **65**, 161–172 (1970).
34. C.R. Rao, Estimation of variance and covariance components: MINQUE theory, *J. of Multivariate Analysis* **1**, 257–277 (1971).
35. S.D. Horn, R.A. Horn and D.B. Duncan, Estimating heteroskedastic variances in linear models, *J. Amer. Stat. Assoc.* **70**, 380–385 (1975).
36. B. Rosenberg, The analysis of a cross section of time series by stochastically convergent parameter regression, *Ann. of Economic and Social Measurement* **2**, 399–428 (1973).
37. A.H. Jazwinski, *Stochastic Processes and Filtering Theory*, Academic Press, New York, (1970).
38. G.C. Chow, *Econometrics*, McGraw-Hill, New York, (1983).
39. W. Schneider, Analytical uses of Kalman filtering in econometrics—A Survey, *Statistical Papers* **29**, 3–33 (1988).
40. J.D. Hamilton, A standard error for the estimated state vector of a state space model when parameters must be estimated from data, *J. of Econometrics* **33**, 387–397 (1986).
41. A. Gelb (ed.), *Applied Optimal Estimation*, M.I.T. Press, Cambridge, Massachusetts, (1974).